Japanese Named Entity Extraction Evaluation - Analysis of Results -

Satoshi Sekine
Computer Science Department
New York University
715 Broadway, 7th floor
New York, NY 10003, USA
sekine@cs.nyu.edu

Yoshio Eriguchi
Research and Development Headquarters
NTT Data Corporation
1-21-2, Shinkawa, Chuo-ku,
Tokyo, 104-0033, Japan
eriguchi@rd.nttdata.co.jp

Abstract

We will report on one of the two tasks in the IREX (Information Retrieval and Extraction Exercise) project, an evaluation-based project for Information Retrieval and Information Extraction in Japanese (Sekine and Isahara, 2000) (IREX Committee, 1999). The project started in 1998 and concluded in September 1999 with many participants and collaborators (45 groups in total from Japan and the US). In this paper, the Named Entity (NE) task is reported. It is a task to extract NE's, such as names of organizations, persons, locations and artifacts, time expressions and numeric expressions from newspaper articles. First, we will explain the task and the definition, as well as the data we created and the results. Second, the analyses of the results will be described, which include analysis of task difficulty across the NE types and system types, analysis of domain dependency and comparison to human performance.

1 Introduction

The need for IR and IE technologies is getting larger because of the improvements in computer technology and the appearance of the Internet. Many researchers in the field feel that the evaluation based projects in the USA, MUC (MUC Homepage, 1999) and TREC (TREC Homepage, 2000), have played a very important role in each field. In Japan, however, while there has been good research, we have had some difficulties comparing systems based on the same platform, since our research is conducted at many different universities, companies, and laboratories using different data and evaluation measures. Our goal is to have a common platform in order to evaluate systems with the same standard. We believe such projects are useful not only for comparing system performance but also to address the following issues:

1) To share and exchange problems among researchers,
2) To accumulate large quantities of data,
3) To let other people know the importance and the quality of IR and IE techniques.

Finishing the project, we believe we achieved these goals.

In this paper we will describe one of the two tasks in the IREX project, the Named Entity task.

2 IREX NE

2.1 Task

Named Entity extraction involves finding Named Entities, such as names of organizations, persons, locations, and artifacts, time expressions, and numeric expressions, such as money and percentage expressions. It is one of the basic techniques used in IR and IE. At the evaluation, participants were asked to identify NE expressions as correctly as possible. In order to avoid a copyright problem, we made a tool to convert a tagged text to a set of tag offset information and we only exchanged tag offset information.

2.2 Definition

The definition of NE's is given in an 18-page document , which is available through the IREX homepage (IREX Homepage, 1999). There are 8 kinds of NE's shown in Table 1. In order to avoid requiring a unique decision for ambiguous cases where even a human could not tag unambiguously, we introduced a tag "OPTIONAL"[1]. If a system tags an expression

[1]This tag is newly introduced in IREX and does not exist in MUC. The tag accounts for 5.7% of all NE occur-

NE	Example
ORGANIZATION	The Diet, IREX Committee
PERSON	Sekine, Wakanohana
LOCATION	Japan, Tokyo, Mt.Fuji
ARTIFACT	Pentium II, Nobel Prize
DATE	March 5, 1965; Yesterday
TIME	11 PM, midnight
MONEY	100 yen, $12,345
PERCENT	10%, a half

Table 1: NE Classes

within the OPTIONAL tag, it is just ignored for the scoring. The definition was created based on the MUC/MET definition; however, the process of making the definition was not easy. In particular, the definition of the newly introduced NE type "artifact" was controversial. We admit that more consideration is needed to make a clearer definition of the NE types.

Comparing the NE task in Japanese to that in English, one of the difficulties comes from the fact that there is no word delimiter in Japanese. Systems have to identify the boundaries of expressions. This will become complicated when we want to tag a substring of what is generally considered a Japanese word. For example, in Japanese there is a word "Rainichi" which means "Visit Japan" and consists of two Chinese characters, "Rai" (Visit) and "Nichi" (abbreviation of Japan). Although many word segmenters identify it as a single word, we expect to extract only "Nichi" as a location. This is a tricky problem, as opposed to the case in English where a word is the unit of NE candidates.

2.3 Runs and Data

There were three kinds of NE exercises, the dry run, a restricted domain formal run, and a general domain formal run, which will be explained later. Also we created three kinds of training data: the dry run training data, the CRL_NE data and the formal run domain restricted training data. Table 2 shows the size of each data set. Note that CRL_NE data belongs to the Communication Research Laboratory (CRL), but it is included in the table, because the data was created by IREX participants, using the definition of IREX-NE, and distributed through IREX.

Data	Number of articles
Dry Run training	46
Dry Run	36
CRL_NE data	1174
Formal run (restricted) training	23
Formal run (restricted)	20
Formal run (general)	71

Table 2: Data size

In order to ensure the fairness of the exercise in the formal run, we used newspaper articles which no one had ever seen. We set the date to freeze the system development (April 13, 1999). The date for the evaluation was set one month after that date (May 13 to 17, 1999) so that we could select the test articles from the period between those dates. We thank the Mainichi Newspaper Corporation for providing this data for us free of charge.

2.4 Restricted domain

In the formal run, in order to study system portability and the effect of domains on NE performance, we had two kinds of evaluation: restricted domain and general domain. In the general domain evaluation, we selected articles regardless of domain. The domain of the restricted domain evaluation was announced one month before the development freeze date. It was an "arrest" domain defined as follows and all the articles in the restricted domain are selected based on the definition.

> *The articles are related to an event "arrest". The event is defined as the arrest of a suspect or suspects by police, National Police, State police or other police forces including the ones of foreign countries. It includes articles mentioning an arrest event in the past. It excludes articles which have only information about requesting an arrest warrant, an accusation or sending the papers pertaining to a case to an Attorney's Office.*

rences in the general domain evaluation and 2.1% in the restricted domain evaluation (the types of the evaluation will be explained later).

2.5 Results

8 groups and 11 systems participated in the dry run, and 14 groups and 15 systems participated in the formal run[2]. The evaluation results were made public anonymously using system ID's. Table 3 shows the evaluation results (F-measure) of the formal run. F-measure is calculated from recall and precision (IREX Committee, 1999). It ranges from 0 to 100, and the larger the better

System ID	general	restrict	diff.
1201	57.69	54.17	-3.52
1205	80.05	78.08	-1.97
1213	66.60	59.87	-6.73
1214	70.34	80.37	**+10.03**
1215	66.74	74.56	+7.82
1223	72.18	74.90	+2.72
1224	75.30	77.61	+2.31
1227	77.37	85.02	+7.65
1229	57.63	64.81	+7.18
1231	74.82	81.94	**+7.12**
1234	71.96	72.77	+0.81
1240	60.96	58.46	-2.50
1247	83.86	87.43	+3.57
1250a	69.82	70.12	+0.30
1250b	57.76	55.24	-2.52

Table 3: NE Formal run result

3 Analyses of the results

3.1 Difficulty across NE type

In Table 4, the F-measure of the best performing system is shown in the "Best" column; the average F-measures are shown in the "Average" column for each NE type on the formal runs. It can be observed that identifying time and numeric expressions is relatively easy, as the average F-measures are more than 80%. In contrast, the accuracy of the other types of NE is not so good. In particular, artifacts are quite difficult to identify. It is interesting to see that tagging artifacts in the general domain is much harder than in the restricted domain. This is because of the limited types of artifacts in the restricted domain. Most of the artifacts in the restricted domain are the names of laws, as the domain is the arrest domain. Systems might be able to find such types of names easily because they could be recognized by a small number of simple patterns or by a short list. The types of the artifacts in the general domain are quite diverse, including names of prizes, novels, ships, or paintings. It might be difficult to build patterns for these items, or systems may need very complicated rules or large dictionaries.

[2]The participation to the dry run was not obligatory. This is why the number of participants is smaller in the dry run than that in the formal run.

3.2 Three types of systems

Based on the questionnaire for the participants we gathered after the formal runs, we found that there are three types of systems.

- Hand created pattern based
  These are pattern based systems where the patterns are created by hand. A typical system used prefix, suffix and proper noun dictionaries. Patterns in these systems look like "If proper nouns are followed by a suffix of person name (for example, a common suffix like "`San`", which is almost equivalent to Mr. and Ms.) then the proper nouns are a person name". This type of system was very common; there were 8 systems in this category.
- Automatically created pattern based
  These are pattern based systems where some or all of the patterns are created automatically using a training corpus. There were three systems in this category, and these systems used quite different methods. One of them used the "error driven method", in which hand created patterns were applied to tagged training data and the system learned from the mistakes. Another system learned patterns for a wide range of information, including, syntax, verb frame and discourse information from training data. The last system used the local context of training data and several filters were applied to get more accurate patterns.
- Fully automatic
  Systems in this category created their knowledge automatically from a training corpus. There were four systems in this category. These systems basically tried to assign one of the four tags, beginning, middle or ending of an NE, or out-of-NE, to

NE type	General domain			Restrict domain			
	Best	Average	Expert	Best	Average	Novice	Expert
Organization	78	57	96	75	55	88	98
Person	87	68	99	87	69	97	100
Location	84	70	98	88	68	94	99
Artifact	44	26	90	83	58	74	92
Date	90	86	98	93	89	96	100
Time	82	83	97	97	90	98	98
Money	86	86	100	100	91	100	100
Percent	84	86	97	-	-	-	-
Total	84	70	98	87	72	94	99

Table 4: Results

each word or each character. The source information for the training was typically character type, POS, dictionary information or lexical information. As the learning mechanism, Maximum Entropy models, decision trees, and HMMs were used.

It is interesting to see that the top three systems came from each category; the best system was a hand created pattern based system, the second system was an automatically created pattern based system and the third system was a fully automatic system. So we believe we can not conclude which type is superior to the others.

Analyzing the results of the top three systems, we observed the importance of the dictionaries. The best hand created pattern based system seems to have a wide coverage dictionary for person, organization and location names and achieved very good accuracy for those categories. However, the hand created pattern based system failed to capture the evaluation specific patterns like "the middle of April". Systems were required to extract the entire expression as a date expression, but the system only extracted "April". The best hand created rule based system, as well as the best automatically created pattern based system also missed other specific patterns which include abbreviations ("`Rai-Nichi`" = Visit-Japan), conjunctions of locations ("`Nichi-Bei`" = Japan-US), and street addresses ("`Meguro-ku, Ookayama 2-12-1`"). The best fully automatic system was successful in extracting most of these specific patterns. However, the fully automatic system has a problem in its coverage. In particular, the training data was newspaper articles published in 1994 and the test data was from 1999, so there are several new names, e.g. the prime minister's name which is not so common (`Obuchi`) and a location name like "Kosovo", which were rarely mentioned in 1994 but appeared a lot in 1999. The system missed many of them.

3.3 Domain dependency

In Table 3, the differences in performance between the general domain and the restricted domain are shown in the column "diff.". Many systems performed better in the restricted domain, although a small number of systems performed better in the general domain. There were two systems which intentionally tuned their systems towards the restricted domain, which are shown in bold in the table. Both of these were among the systems which performed much better (more than 7%) in the restricted domain. The system which achieved the largest improvement was a fully automatic system, and it only replaced the training data for the domain restricted task (so this is an intentionally tuned system). It shows the domain dependency of the task, although further investigation is needed to see why some other systems can perform much better even without domain tuning.

3.4 Comparison to human performance

In Table 4, human performance is shown in the "Novice" and "Expert" columns. "Novice" means the average F-measure of three graduate students and "Expert" means the average F-measure of the two people who were most re-

sponsible for creating the definition and created the answer. They first created two answers independently and checked them by themselves. The results after the checking are shown in the table, so many careless mistakes were deleted at this time. We can say that 98-99 F-measure is the performance of experts who create them very carefully, and 94 is a usual person's performance.

We can find a similar pattern of performance among different NEs. Humans also performed more poorly for artifacts and very well for time and numeric expressions.

The difference between the best system performance and human performance is 7 or more F-measure, as opposed to the case in English where the top systems perform at a level comparable or superior to human performance. There could be several reasons for this. One obvious reason is that we introduced a difficult NE type, artifact, which degrades the overall performance more for the system side than the human side. Also, the difficulty of identifying the expression boundaries may contribute to the difference. Finally, we believe that the systems can possibly improve, as IREX was the first evaluation based project in Japanese, whereas in English there have been 7 MUC's and the technology may have matured by now.

4 Conclusion

We reported on the NE task of the IREX project. We first explained the task and the definition, as well as the data we created and the results. The analyses of the result were described, which include analysis of task difficulty across the NE types and system types, analysis of domain dependency and comparison to human performance.

As this is one of the first projects of this type in Japan, we may have a lot to do in the future and hopefully the results of the project will be beneficial for future projects. As the next step, IREX will be merged with a similar project NTCIR (NTCIR Homepage, 2000) which places more emphasis on IR, with a newly created project for summarization, TSC (TSC Homepage, 2000), and continue this kind of effort for the future.

5 Acknowledgment

We would like to thank all the participants of IREX projects. The project would never have been as successful as it was without the participants, all of whom were very cooperative and constructive.

References

IREX Committee 1999 *Proceedings of the IREX Workshop*

Satoshi Sekine, Hitoshi Isahara. 2000 : "IREX: IR and IE Evaluation Project in Japanese" *Proceedings of the LREC-2000 conference*

IREX Homepage http://cs.nyu.edu/projects/proteus/irex

MUC Homepage http://www.muc.saic.com/

TREC Homepage trec.nist.gov/

NTCIR Homepage http://www.rd.nacsis.ac.jp/ñtcadm/index-en.html

TSC Homepage http://galaga.jaist.ac.jp:8000/tsc/